GeDi: Generative Discriminator Guided Sequence Generation

Year: 2020 | Citations: 450 | Authors: Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, N. Keskar, Shafiq R. Joty

---

Abstract

While large-scale language models (LMs) are able to imitate the distribution of natural language well enough to generate realistic text, it is difficult to control which regions of the distribution they generate. This is especially problematic because datasets used for training large LMs usually contain significant toxicity, hate, bias, and negativity. We propose GeDi as an efficient method for using smaller LMs as generative discriminators to guide generation from large LMs to make them safer and more controllable. GeDi guides generation at each step by computing classification probabilities for all possible next tokens via Bayes rule by normalizing over two class-conditional distributions; one conditioned on the desired attribute, or control code, and another conditioned on the undesired attribute, or anti control code. We find that GeDi gives stronger controllability than the state of the art method while also achieving generation speeds more than 30 times faster. Additionally, training GeDi on only four topics allows us to controllably generate new topics zero-shot from just a keyword, unlocking a new capability that previous controllable generation methods do not have. Lastly, we show that GeDi can make GPT-2 (1.5B parameters) significantly less toxic without sacrificing linguistic quality, making it by far the most practical existing method for detoxifying large language models while maintaining a fast generation speed.